Testing the significance of assuming homogeneity in contingency-tables/cross-tabulations

Mark Tygert

January 9, 2012

Abstract

The model for homogeneity of proportions in a two-way contingency-table/cross-tabulation is the same as the model of independence, except that the probabilistic process generating the data is viewed as fixing the column totals (but not the row totals). When gauging the consistency of observed data with the assumption of independence, recent work has illustrated that the Euclidean/Frobenius/Hilbert-Schmidt distance is often far more statistically powerful than the classical statistics such as χ^2, the log–likelihood-ratio G^2, the Freeman-Tukey/Hellinger distance, and other members of the Cressie-Read power-divergence family. The present paper indicates that the Euclidean/Frobenius/Hilbert-Schmidt distance can be more powerful for gauging the consistency of observed data with the assumption of homogeneity, too.

Keywords: chi-square, Fisher's exact, Freeman-Tukey, likelihood ratio, power divergence, root-mean-square

1 Introduction

The statistical analysis of categorical data is commonly formulated in the framework of contingency-tables/cross-tabulations; Table 1 provides a typical two-way example (see, for instance, Chapter 4 of Andersen, 1990, for a comprehensive treatment). A common task is to ascertain whether the given data (displayed in Table 1) is consistent up to expected statistical fluctuations with the model for homogeneity of proportions (displayed in Table 2). When considering homogeneity, we assume that the probabilistic process generating the given data fixes the column totals (but not the row totals) by construction. Therefore, to gauge whether the given data displayed in Table 1 is consistent with the assumed homogeneity displayed in Table 2, we do the following:

1. We generate s sets of draws, with the kth set consisting of $n_{\bullet,k}$ independent and identically distributed draws from the probability distribution $(p_1, p_2, \ldots, p_r)$, where $p_1 = n_{1,\bullet}/n$, $p_2 = n_{2,\bullet}/n$, $\ldots$, $p_r = n_{r,\bullet}/n$. Note that $p_j = (n_{j,\bullet} \cdot n_{\bullet,k}/n)/n_{\bullet,k}$ for $j = 1, 2, \ldots, r$; these are homogeneous proportions (since p_1, p_2, $\ldots$, p_r are the same for every column index k).

Table 1: A typical two-way contingency-table/cross-tabulation ($n_{j,k}$ is a nonnegative integer with $j = 1, 2, \ldots, r, \quad k = 1, 2, \ldots, s$; $n_{j,\bullet} = \sum_{k=1}^{s} n_{j,k}$ is a row total with $j = 1, 2, \ldots, r$; $n_{\bullet,k} = \sum_{j=1}^{r} n_{j,k}$ is a column total with $k = 1, 2, \ldots, s$; and $n_{\bullet,\bullet} = \sum_{j=1}^{r} \sum_{k=1}^{s} n_{j,k} = n$ is the grand total)

	1	2	$\cdots$	s	
1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,s}$	$n_{1,\bullet}$
2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,s}$	$n_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$n_{r,1}$	$n_{r,2}$	$\cdots$	$n_{r,s}$	$n_{r,\bullet}$
	$n_{\bullet,1}$	$n_{\bullet,2}$	$\cdots$	$n_{\bullet,s}$	$n_{\bullet,\bullet}$

Table 2: The model for homogeneity of proportions ($n_{1,\bullet}$, $n_{2,\bullet}$, $\ldots$, $n_{r,\bullet}$ are the row totals; $n_{\bullet,1}$, $n_{\bullet,2}$, $\ldots$, $n_{\bullet,s}$ are the column totals; and $n_{\bullet,\bullet} = n$ is the grand total)

	1	2	$\cdots$	s	
1	$n_{1,\bullet} \cdot n_{\bullet,1}/n$	$n_{1,\bullet} \cdot n_{\bullet,2}/n$	$\cdots$	$n_{1,\bullet} \cdot n_{\bullet,s}/n$	$n_{1,\bullet}$
2	$n_{2,\bullet} \cdot n_{\bullet,1}/n$	$n_{2,\bullet} \cdot n_{\bullet,2}/n$	$\cdots$	$n_{2,\bullet} \cdot n_{\bullet,s}/n$	$n_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	$n_{r,\bullet} \cdot n_{\bullet,1}/n$	$n_{r,\bullet} \cdot n_{\bullet,2}/n$	$\cdots$	$n_{r,\bullet} \cdot n_{\bullet,s}/n$	$n_{r,\bullet}$
	$n_{\bullet,1}$	$n_{\bullet,2}$	$\cdots$	$n_{\bullet,s}$	$n_{\bullet,\bullet}$

2. For each of the s sets of draws — say the kth set — we define $N_{j,k}$ to be the number of draws falling in the jth row, for $j = 1, 2, \ldots, r$.

3. We calculate the probability P that the discrepancy between the simulated counts $N_{j,k}$ and the model $N_{j,\bullet} \cdot N_{\bullet,k}/n$ is greater than or equal to the discrepancy between the observed counts $n_{j,k}$ and the assumed $n_{j,\bullet} \cdot n_{\bullet,k}/n$. When calculating this probability, we view $N_{j,k}$ and $N_{j,\bullet}$ as random, while viewing all other numbers as fixed. Please note that, by construction, $N_{\bullet,k} = n_{\bullet,k}$ for $k = 1, 2, \ldots, s$.

The number P defined in Step 3 is known as the (exact) P-value. Given the P-value P, we can have $100(1 - P)\%$ confidence that the observed draws are not consistent with assuming the homogeneity displayed in Table 2. See Section 3 of Perkins et al. (2011) for further discussion of P-values and their interpretation; Section 3 of Perkins et al. (2011) details subtleties involved in the definition and interpretation of these P-values.

The definition above of the P-value P requires a metric for measuring the discrepancies. The canonical choices are χ^2 and the log–likelihood-ratio G^2:

$$\chi^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(n_{j,k} - (n_{j,\bullet} \cdot n_{\bullet,k}/n))^2}{n_{j,\bullet} \cdot n_{\bullet,k}/n} \tag{1}$$

$$X^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(N_{j,k} - (N_{j,\bullet} \cdot N_{\bullet,k}/n))^2}{N_{j,\bullet} \cdot N_{\bullet,k}/n} \tag{2}$$

$$P_{\chi^2} = \text{Prob}\{X^2 \geq \chi^2\} \tag{3}$$

$$g^2 = 2 \sum_{j=1}^{r} \sum_{k=1}^{s} n_{j,k} \cdot \ln\left(\frac{n_{j,k}}{n_{j,\bullet} \cdot n_{\bullet,k}/n}\right) \tag{4}$$

$$G^2 = 2 \sum_{j=1}^{r} \sum_{k=1}^{s} N_{j,k} \cdot \ln\left(\frac{N_{j,k}}{N_{j,\bullet} \cdot N_{\bullet,k}/n}\right) \tag{5}$$

$$P_{g^2} = \text{Prob}\{G^2 \geq g^2\} \tag{6}$$

Other possibilities include the Hellinger (or Freeman-Tukey) distance and the Frobenius (or Hilbert-Schmidt or Euclidean) distance:

$$h^2 = 4 \sum_{j=1}^{r} \sum_{k=1}^{s} (\sqrt{n_{j,k}} - \sqrt{n_{j,\bullet} \cdot n_{\bullet,k}/n})^2 \tag{7}$$

$$H^2 = 4 \sum_{j=1}^{r} \sum_{k=1}^{s} (\sqrt{N_{j,k}} - \sqrt{N_{j,\bullet} \cdot N_{\bullet,k}/n})^2 \tag{8}$$

$$P_{h^2} = \text{Prob}\{H^2 \geq h^2\} \tag{9}$$

$$f^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} (n_{j,k} - (n_{j,\bullet} \cdot n_{\bullet,k}/n))^2 \tag{10}$$

$$F^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} (N_{j,k} - (N_{j,\bullet} \cdot N_{\bullet,k}/n))^2 \tag{11}$$

$$P_{f^2} = \text{Prob}\{F^2 \geq f^2\} \tag{12}$$

When taking probabilities in (3), (6), (9), and (12), we view the uppercase X^2, G^2, H^2, and F^2 as random variables, while viewing the lowercase χ^2, g^2, h^2, and f^2 as fixed numbers.

As discussed, for example, by Rao (2002), X^2, G^2, and H^2 all converge to the same distribution in the limit of large numbers of draws — X^2, G^2, and H^2 are the best-known members of the Cressie-Read power-divergence family. F^2 is not a member of the Cressie-Read power-divergence family and does not necessarily converge to the same distribution as X^2, G^2, and H^2. Perkins et al. (2011) illustrated the many advantages of F^2 when neither the row totals nor the column totals are fixed; the present paper illustrates the advantages when the column totals are fixed. However, F^2 is not uniformly more powerful than the classical statistics. We recommend using both F^2 and a classical statistic such as G^2.

In the sequel, Section 2 summarizes an algorithm for computing the P-values defined above. Section 3 analyzes several data sets. Section 4 draws some conclusions.

2 Computation of P-values

The definitions of the P-values in (3), (6), (9), and (12) involve the probabilities of certain events. In the present paper, we compute these probabilities via Monte-Carlo simulations with guaranteed error bounds. Specifically, we conduct a large number m of simulations; in each simulation — say the ℓth — we perform the following steps (using the data of Table 1):

1. We generate s sets of draws, with the kth set consisting of $n_{\bullet,k}$ independent and identically distributed draws from the probability distribution $(p_1, p_2, \ldots, p_r)$, where $p_1 = n_{1,\bullet}/n$, $p_2 = n_{2,\bullet}/n$, ..., $p_r = n_{r,\bullet}/n$. Note that $p_j = (n_{j,\bullet} \cdot n_{\bullet,k}/n)/n_{\bullet,k}$ for $j = 1, 2, \ldots, r$; these are homogeneous proportions (since p_1, p_2, ..., p_r are the same for every column index k). Furthermore, the underlying distribution of the draws does not depend on ℓ.

2. For each of the s sets of draws — say the kth set — we define $n_{j,k}^{(\ell)}$ to be the number of draws falling in the jth row, for $j = 1, 2, \ldots, r$.

3. We calculate the discrepancy $f_{(\ell)}^2$ between the simulated counts $n_{j,k}^{(\ell)}$ and the model $n_{j,\bullet}^{(\ell)} \cdot n_{\bullet,k}^{(\ell)}/n$, that is,
$$f_{(\ell)}^2 = \sum_{j=1}^{r} \sum_{k=1}^{s} (n_{j,k}^{(\ell)} - (n_{j,\bullet}^{(\ell)} \cdot n_{\bullet,k}^{(\ell)}/n))^2. \tag{13}$$

An estimate of the P-value P_{f^2} is the fraction of $f_{(1)}^2$, $f_{(2)}^2$, ..., $f_{(m)}^2$ which are greater than or equal to f^2 defined in (10). As discussed in Section 3 of Perkins et al. (2011), the standard error of the estimate is $\sqrt{P_{f^2}(1 - P_{f^2})/m}$, where m is the number of simulations.

Needless to say, we can compute the P-values for χ^2, g^2, and h^2 via similar procedures, with the same error bounds.

Remark 2.1. For all computations reported in the present paper, we generated random numbers via the C programming language procedure given on page 9 of Marsaglia (2003), implementing the recommended complementary multiply with carry.

3 Data analysis

To compare the performance of the various metrics for measuring the discrepancies between observed and simulated data, we analyze several data sets. Using the procedure of Section 2, we conduct m = 4,000,000 Monte-Carlo simulations per P-value, for each of the examples presented below. The standard error of the resulting estimate for the P-value P is then $\sqrt{P(1 - P)}/2000$; see Section 3 of Perkins et al. (2011). Before reporting the P-values associated with the data sets, we make two remarks concerning their interpretation:

Remark 3.1. A significance test can only indicate that observed data *cannot* be reasonably assumed to have arisen from the model of homogeneous proportions; a significance test cannot prove that the observed data *can* be reasonably assumed to have arisen from the model of homogeneity. Thus, aside from considerations of multiple testing, if any statistic

strongly signals that the data cannot be reasonably assumed to have arisen from the model of homogeneity, then we must reject (or at least question) the model — irrespective of any large P-values for other statistics. For instance, if the P-value for the Frobenius distance f^2 is very small, then we should not accept the model of homogeneity, not even if the P-values for χ^2, the log–likelihood-ratio g^2, and the Freeman-Tukey/Hellinger-distance h^2 are large.

Remark 3.2. The term "negative log-likelihood" used in the present section refers to the statistic that is simply the negative of the logarithm of the likelihood. The negative log-likelihood is the same statistic used in the generalization of Fisher's exact test discussed by Guo and Thompson (1992); unlike the log–likelihood-ratio G^2, this statistic involves only one likelihood, not the ratio of two. We mention the negative log-likelihood just to facilitate comparisons; we are not asserting that the likelihood on its own (rather than in a ratio) is a good gauge of the relative sizes of deviations from a model.

The 11×2 Table 3 displays the data for our first example, which has 22 entries in all. Table 4 displays the model of homogeneous proportions for Table 3. The P-values for Table 3 for the assumption that Table 4 gives the correct underlying distribution are

χ^2 (X^2):	.0868
log–likelihood-ratio (G^2):	.0906
Freeman-Tukey/Hellinger (H^2):	.0959
negative log-likelihood:	.0905
Frobenius (F^2):	.00838

Please note that the P-value for the Frobenius distance is over an order of magnitude smaller than the P-values for the classical statistics.

The 7×3 Table 7 displays the data for our second example, which has 21 entries in all. Table 8 displays the model of homogeneous proportions for Table 7. The P-values for Table 7 for the assumption that Table 8 gives the correct underlying distribution are

χ^2 (X^2):	.145
log–likelihood-ratio (G^2):	.292
Freeman-Tukey/Hellinger (H^2):	.493
negative log-likelihood:	.132
Frobenius (F^2):	.0286

Please note that the P-value for the Frobenius distance is over four times smaller than the P-values for the classical statistics.

The 9×2 Table 11 displays the data for our third example, which has 18 entries in all. Table 12 displays the model of homogeneous proportions for Table 11. The P-values for Table 11 for the assumption that Table 12 gives the correct underlying distribution are

χ^2 (X^2):	.123
log–likelihood-ratio (G^2):	.138
Freeman-Tukey/Hellinger (H^2):	.157
negative log-likelihood:	.114
Frobenius (F^2):	.0344

Please note that the P-value for the Frobenius distance is over three times smaller than the P-values for the classical statistics.

The 5×3 Table 15 displays the data for our final example, which has 15 entries in all. Table 16 displays the model of homogeneous proportions for Table 15. The P-values for Table 15 for the assumption that Table 16 gives the correct underlying distribution are

χ^2 (X^2):	.276
log–likelihood-ratio (G^2):	.171
Freeman-Tukey/Hellinger (H^2):	.0794
negative log-likelihood:	.235
Frobenius (F^2):	.199

In this example, none of the statistics produces a very small P-value; the smallest arises from the Freeman-Tukey/Hellinger distance in this case.

Remark 3.3. Appropriate binning (or rebinning) to uniformize the frequencies associated with the entries in the contingency-tables/cross-tabulations can mitigate the problem with the classical statistics. Yet rebinning is a black art that is liable to improperly influence the result of a significance test, and the usual data-dependent rebinning calls for Monte-Carlo simulations to calculate P-values accurately anyways. Rebinning always requires careful extra work. A principal advantage of the Frobenius distance is that it does not require any rebinning; indeed, the Frobenius distance is most powerful without any rebinning. Note also that optimally rebinning data such as that displayed in Table 3 can be very challenging.

4 Conclusion

The Frobenius distance is significantly more powerful than the classical statistics for gauging the consistency of observed data with the assumption of homogeneity in many of the examples of the present paper. This may or may not be typical of most applications; actually, we suspect that our last example — in which all the statistics perform similarly — is the most representative. Even so, both the present paper and the applications of Perkins et al. (2011) illustrate that there are many important circumstances in which the Frobenius distance is much more powerful than the classical statistics.

Acknowledgements

We would like to thank Alex Barnett, Gérard Ben Arous, James Berger, Tony Cai, Sourav Chatterjee, Ronald Raphael Coifman, Ingrid Daubechies, Jianqing Fan, Jiayang Gao, Andrew Gelman, Leslie Greengard, Peter W. Jones, Deborah Mayo, Peter McCullagh, Michael O'Neil, Ron Peled, William Perkins, William H. Press, Vladimir Rokhlin, Joseph Romano, Gary Simon, Amit Singer, Michael Stein, Stephen Stigler, Joel Tropp, Rachel Ward, Larry Wasserman, and Douglas A. Wolfe. This work was supported in part by a research fellowship from the Alfred P. Sloan Foundation.

Table 3: Results of polls in June 1983 for Danish parliamentary elections, from Chapter 4 of Andersen (1990)

Party	Poll 1		Poll 2	
A	416	(33.1%)	268	(38.9%)
B	45	(3.6%)	22	(3.2%)
C	338	(26.9%)	160	(23.2%)
E	13	(1.0%)	6	(0.9%)
F	131	(10.4%)	66	(9.6%)
K	18	(1.4%)	10	(1.5%)
M	47	(3.7%)	16	(2.3%)
Q	20	(1.6%)	8	(1.2%)
V	129	(10.3%)	92	(13.4%)
Y	22	(1.8%)	9	(1.3%)
Z	76	(6.1%)	32	(4.6%)
All	1255	(100.0%)	689	(100.0%)

Table 4: The model of homogeneous proportions for Table 3

Party	Poll 1		Poll 2	
A	441.6	(35.2%)	242.4	(35.2%)
B	43.3	(3.4%)	23.7	(3.4%)
C	321.5	(25.6%)	176.5	(25.6%)
E	12.3	(1.0%)	6.7	(1.0%)
F	127.2	(10.1%)	69.8	(10.1%)
K	18.1	(1.4%)	9.9	(1.4%)
M	40.7	(3.2%)	22.3	(3.2%)
Q	18.1	(1.4%)	9.9	(1.4%)
V	142.7	(11.4%)	78.3	(11.4%)
Y	20.0	(1.6%)	11.0	(1.6%)
Z	69.7	(5.6%)	38.3	(5.6%)
All	1255.0	(100.0%)	689.0	(100.0%)

Table 5: Differences between the entries of Table 3 and the corresponding entries of Table 4

Party	Poll 1	Poll 2
A	−25.6	25.6
B	1.7	−1.7
C	16.5	−16.5
E	0.7	−0.7
F	3.8	−3.8
K	−0.1	0.1
M	6.3	−6.3
Q	1.9	−1.9
V	−13.7	13.7
Y	2.0	−2.0
Z	6.3	−6.3
All	0.0	0.0

Table 6: The entries of Table 5 divided by the square roots of the corresponding entries of Table 4

Party	Poll 1	Poll 2
A	−1.2	1.6
B	0.3	−0.4
C	0.9	−1.2
E	0.2	−0.3
F	0.3	−0.5
K	−0.0	0.0
M	1.0	−1.3
Q	0.5	−0.6
V	−1.1	1.5
Y	0.4	−0.6
Z	0.8	−1.0
All	0.0	0.0

Table 7: Reasons for (or absence of) premature termination of the treatment of maniacal patients in three groups from Bowden et al. (1994) (the three groups are those treated with divalproex, those treated with lithium, and those "treated" with a placebo)

Reason	Divalproex		Lithium		Placebo	
Lack of efficacy	21	(30.4%)	12	(33.3%)	38	(51.4%)
Intolerance	4	(5.8%)	4	(11.1%)	2	(2.7%)
Recovered	3	(4.3%)	2	(5.6%)	2	(2.7%)
Noncompliance	1	(1.4%)	1	(2.8%)	3	(4.1%)
Another illness	0	(0.0%)	1	(2.8%)	0	(0.0%)
Administration	4	(5.8%)	2	(5.6%)	2	(2.7%)
Not terminated	36	(52.2%)	14	(38.9%)	27	(36.5%)
All	69	(100.0%)	36	(100.0%)	74	(100.0%)

Table 8: The model of homogeneous proportions for Table 7

Reason	Divalproex		Lithium		Placebo	
Lack of efficacy	27.4	(39.7%)	14.3	(39.7%)	29.4	(39.7%)
Intolerance	3.9	(5.6%)	2.0	(5.6%)	4.1	(5.6%)
Recovered	2.7	(3.9%)	1.4	(3.9%)	2.9	(3.9%)
Noncompliance	1.9	(2.8%)	1.0	(2.8%)	2.1	(2.8%)
Another illness	0.4	(0.6%)	0.2	(0.6%)	0.4	(0.6%)
Administration	3.1	(4.5%)	1.6	(4.5%)	3.3	(4.5%)
Not terminated	29.7	(43.0%)	15.5	(43.0%)	31.8	(43.0%)
All	69.0	(100.0%)	36.0	(100.0%)	74.0	(100.0%)

Table 9: Differences between the entries of Table 7 and the corresponding entries of Table 8

Reason	Divalproex	Lithium	Placebo
Lack of efficacy	−6.4	−2.3	8.6
Intolerance	0.1	2.0	−2.1
Recovered	0.3	0.6	−0.9
Noncompliance	−0.9	0.0	0.9
Another illness	−0.4	0.8	−0.4
Administration	0.9	0.4	−1.3
Not terminated	6.3	−1.5	−4.8
All	0.0	0.0	0.0

Table 10: The entries of Table 9 divided by the square roots of the corresponding entries of Table 8

Reason	Divalproex	Lithium	Placebo
Lack of efficacy	−1.2	−0.6	1.6
Intolerance	0.1	1.4	−1.0
Recovered	0.2	0.5	−0.5
Noncompliance	−0.7	0.0	0.6
Another illness	−0.6	1.8	−0.6
Administration	0.5	0.3	−0.7
Not terminated	1.2	−0.4	−0.9
All	0.0	0.0	0.0

Table 11: Results for the 2012 Republican U.S. presidential nomination, from a CBS News poll of November 6–10, 2011 (released November 11, 2011) and from a Pew Research Center poll of November 9–11, 2011 (released November 17, 2011), as reconstructed from percentages rounded to the nearest whole numbers (the original counts were not reported) for Republican primary voters

Candidate	CBS		Pew	
Michele Bachmann	15	(4.6%)	21	(5.1%)
Herman Cain	69	(21.2%)	103	(25.0%)
Newt Gingrich	57	(17.5%)	66	(16.0%)
Jon Huntsman	4	(1.2%)	4	(1.0%)
Ron Paul	19	(5.8%)	33	(8.0%)
Rick Perry	31	(9.5%)	37	(9.0%)
Mitt Romney	57	(17.5%)	91	(22.1%)
Rick Santorum	8	(2.5%)	8	(1.9%)
Do not know	65	(20.0%)	49	(11.9%)
All	325	(100.0%)	412	(100.0%)

Table 12: The model of homogeneous proportions for Table 11

Candidate	CBS		Pew	
Michele Bachmann	15.9	(4.9%)	20.1	(4.9%)
Herman Cain	75.8	(23.3%)	96.2	(23.3%)
Newt Gingrich	54.2	(16.7%)	68.8	(16.7%)
Jon Huntsman	3.5	(1.1%)	4.5	(1.1%)
Ron Paul	22.9	(7.1%)	29.1	(7.1%)
Rick Perry	30.0	(9.2%)	38.0	(9.2%)
Mitt Romney	65.3	(20.1%)	82.7	(20.1%)
Rick Santorum	7.1	(2.2%)	8.9	(2.2%)
Do not know	50.3	(15.5%)	63.7	(15.5%)
All	325.0	(100.0%)	412.0	(100.0%)

Table 13: Differences between the entries of Table 11 and the corresponding entries of Table 12

Candidate	CBS	Pew
Michele Bachmann	−0.9	0.9
Herman Cain	−6.8	6.8
Newt Gingrich	2.8	−2.8
Jon Huntsman	0.5	−0.5
Ron Paul	−3.9	3.9
Rick Perry	1.0	−1.0
Mitt Romney	−8.3	8.3
Rick Santorum	0.9	−0.9
Do not know	14.7	−14.7
All	0.0	0.0

Table 14: The entries of Table 13 divided by the square roots of the corresponding entries of Table 12

Candidate	CBS	Pew
Michele Bachmann	−0.2	0.2
Herman Cain	−0.8	0.7
Newt Gingrich	0.4	−0.3
Jon Huntsman	0.3	−0.2
Ron Paul	−0.8	0.7
Rick Perry	0.2	−0.2
Mitt Romney	−1.0	0.9
Rick Santorum	0.4	−0.3
Do not know	2.1	−1.8
All	0.0	0.0

Table 15: Reactions to prior treatment with lithium (when treated before with lithium) of maniacal patients in three groups from Bowden et al. (1994) (the three groups are those treated with divalproex, those treated with lithium, and those "treated" with a placebo)

Reaction	Divalproex		Lithium		Placebo	
Effective and tolerated	22	(31.9%)	16	(44.4%)	19	(25.7%)
Effective but not tolerated	7	(10.1%)	0	(0.0%)	6	(8.1%)
Ineffective but tolerated	19	(27.5%)	11	(30.6%)	31	(41.9%)
Ineffective and not tolerated	6	(8.7%)	4	(11.1%)	5	(6.8%)
No prior lithium treatment	15	(21.7%)	5	(13.9%)	13	(17.6%)
All	69	(100.0%)	36	(100.0%)	74	(100.0%)

Table 16: The model of homogeneous proportions for Table 15

Reaction	Divalproex		Lithium		Placebo	
Effective and tolerated	22.0	(31.8%)	11.5	(31.8%)	23.6	(31.8%)
Effective but not tolerated	5.0	(7.3%)	2.6	(7.3%)	5.4	(7.3%)
Ineffective but tolerated	23.5	(34.1%)	12.3	(34.1%)	25.2	(34.1%)
Ineffective and not tolerated	5.8	(8.4%)	3.0	(8.4%)	6.2	(8.4%)
No prior lithium treatment	12.7	(18.4%)	6.6	(18.4%)	13.6	(18.4%)
All	69.0	(100.0%)	36.0	(100.0%)	74.0	(100.0%)

Table 17: Differences between the entries of Table 15 and the corresponding entries of Table 16

Reaction	Divalproex	Lithium	Placebo
Effective and tolerated	0.0	4.5	−4.6
Effective but not tolerated	2.0	−2.6	0.6
Ineffective but tolerated	−4.5	−1.3	5.8
Ineffective and not tolerated	0.2	1.0	−1.2
No prior lithium treatment	2.3	−1.6	−0.6
All	0.0	0.0	0.0

Table 18: The entries of Table 17 divided by the square roots of the corresponding entries of Table 16

Reaction	Divalproex	Lithium	Placebo
Effective and tolerated	0.0	1.3	−0.9
Effective but not tolerated	0.9	−1.6	0.3
Ineffective but tolerated	−0.9	−0.4	1.2
Ineffective and not tolerated	0.1	0.6	−0.5
No prior lithium treatment	0.6	−0.6	−0.2
All	0.0	0.0	0.0

References

Andersen, E. B. (1990), *The Statistical Analysis of Categorical Data*, Berlin: Springer-Verlag.

Bowden, C. L., Brugger, A. M., Swann, A. C., Calabrese, J. R., Janicak, P. G., Petty, F., Dilsaver, S. C., Davis, J. M., Rush, A. J., Small, J. G., Garza-Treviño, E. S., Risch, S. C., Goodnick, P. J., and Morris, D. D. (1994), Efficacy of divalproex vs. lithium and placebo in the treatment of mania, *J. Amer. Med. Assoc.*, **271**, 918–924.

Guo, S. W. and Thompson, E. A. (1992), Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, **48**, 361–372.

Marsaglia, G. (2003), Random number generators, *J. Modern Appl. Stat. Meth.*, **2**, 2–13.

Perkins, W., Tygert, M., and Ward, R. (2011), χ^2 and classical exact tests often wildly misreport significance; the remedy lies in computers, Tech. Rep. 1108.4126, arXiv, http://cims.nyu.edu/∼tygert/abbreviated.pdf.

Rao, C. R. (2002), Karl Pearson chi-square test: The dawn of statistical inference, In *Goodness-of-Fit Tests and Model Validity* (eds Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., and Mesbah, M.), pp. 9–24, Boston: Birkhäuser.